MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

# LARGE-SCALE NETWORK ANALYSIS FOR ONLINE SOCIAL BRAND ADVERTISING[1]

**Kunpeng Zhang**
Department of DOIT, Robert H. Smith School of Business, University of Maryland,
College Park, MD 27042 U.S.A. {kzhang@rhsmith.umd.edu}

**Siddhartha Bhattacharyya**
Department of IDS, College of Business Administration, University of Illinois, Chicago,
Chicago, IL 60607 U.S.A. {sidb@uic.edu}

**Sudha Ram**
Department of MIS, Eller College of Management, University of Arizona,
Tucson, AZ 85721 U.S.A. {ram@eller.arizona.edu}

*This paper proposes an audience selection framework for online brand advertising based on user activities on social media platforms. It is one of the first studies to our knowledge that develops and analyzes implicit brand–brand networks for online brand advertising. This paper makes several contributions. We first extract and analyze implicit weighted brand–brand networks, representing interactions among users and brands, from a large dataset. We examine network properties and community structures and propose a framework combining text and network analyses to find target audiences. As a part of this framework, we develop a hierarchical community detection algorithm to identify a set of brands that are closely related to a specific brand. This latter brand is referred to as the "focal brand." We also develop a global ranking algorithm to calculate brand influence and select influential brands from this set of closely related brands. This is then combined with sentiment analysis to identify target users from these selected brands. To process large-scale datasets and networks, we implement several MapReduce-based algorithms. Finally, we design a novel evaluation technique to test the effectiveness of our targeting framework. Experiments conducted with Facebook data show that our framework provides significant performance improvements in identifying target audiences for focal brands.*

**Keywords**: Online advertising, brand–brand networks, community detection, audience selection, sentiment analysis

## Introduction

Social media sites such as Facebook, Twitter, and Amazon allow users to generate, share, and communicate with others on topics of interest to them. For example, users may "follow" social brands on Facebook, and comment on, or "like" posts made by a brand page.[2] All these social actions generate a rich data source, which can be analyzed to understand the interactions among entities on social media. Two types of interaction networks may be extracted from these

[1]Bart Baesens, Ravi Bapna, James R. Marsden, Jan Vanthienen, and J. Leon Zhao served as the senior editors for this paper.

The appendices for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

[2]A social brand is any entity such as an institute, an organization, a company, a public celebrity, a service, or a product.

datasets: explicit or implicit. Explicit networks are created via "friendship" relationships on Facebook or "following" relationships on Twitter, while implicit networks are created via users' reviewing actions on products on Amazon.com or through blog subscriptions and comments (Chau and Xu 2012). Implicit networks may also be established when people share common interests, or when brands have overlapping customers. Analysis of such large implicit networks enables the detection of communities of similar users or similar brands, facilitating targeted online advertising that may eventually lead to product or service purchases.

Online advertising is a large and rapidly growing business. Revenues from online advertising in the United States surpassed those of cable television and nearly exceeded those of broadcast television. As indicated in the IAB Internet advertising revenue report in 2014, Internet advertising revenues in the United States totaled $42.8 billion for 2013, an increase of 17% over 2012.[3] Online advertising is a major source of revenue for some leading companies, such as Google, Facebook, and Twitter, and blogging platforms such as Tumblr, Wordpress, and Blogspots.

A major research question in online advertising is to identify target users who are likely to be interested in a specific focal brand. There are many approaches and frameworks deployed by marketing researchers and computer scientists. These include (1) finding a target audience using user profile information (e.g., demographics, geographic); (2) using behavioral data (user historical activities, such as clicking history and purchasing history), if available, to build predictive models to identify potential audiences for advertising; (3) selecting a subset of users who are positively inclined toward the brand via sentiment analysis of their online comments/reviews; and (4) targeting the most influential people (e.g., public celebrities) and their friends or followers via social media analytics. All these approaches are fairly easy to implement; however, they each have several shortcomings. Targeting based on user profile data ignores much of the information on user–user and user–brand interactions. Inaccurate and incomplete user profile data can also reduce targeting performance. Approaches based on text-based sentiment analyses of user-generated content do not make full use of network information. Also, the efficacy of targeting through friendship network influence is not always clear (Van den Bulte 2010; Watts and Dodds 2007), because friends or followers of a user may not necessarily share the same interests in brands.

In this paper, we attempt to overcome these shortcomings by combining network and textual data analysis to develop a new targeting framework for online brand advertising. Unlike regular social advertising that seeks target audiences through users' friend networks (Goldfarb and Tucker 2011), we first extract implicit brand–brand networks from a large amount of historical user–brand interactions to capture relationships among brands. We then develop network analysis algorithms, together with sentiment analysis on user-generated content, to identify target audiences.

The implicit brand–brand network forms the basis for selecting target audiences for a focal brand. We develop a three-step approach to identify the target audience. A potential user for targeting is one who is not currently engaged with the focal brand, but who is interested in other brands that are related/similar to the focal brand. First, related/similar brands are identified through a hierarchical community detection algorithm applied to the undirected brand–brand network. In the second step, a directed network derived from this undirected brand–brand network is used to determine the global influential score for each brand. The results of these two steps are used to generate an initial pool of potential users for targeting. Target audiences are then finally selected from this user pool based on the sentiment of their comments across all brands. The motivating rationale is that users with generally positive sentiment are more likely to propagate positive information about the focal brand.

To evaluate the performance of our targeting framework, we conduct experiments on a large dataset collected from Facebook. Our proposed approach shows significant improvements in terms of audience reach as compared to baselines. Further, the brand–brand networks reveal interesting brand relationships from a consumer's perspective, which will be of broad interest to marketing practitioners and researchers.

Researchers in recent years have noted the potential for leveraging vast troves of online micro-scale data on user activity for deriving actionable insights (Gopal et al. 2011; Marsden 2008). Researchers have also attempted to combine multiple computational techniques to build more intelligent business systems. Analytics approaches on "big" data provide strategic advantage and add value in various business problems and domains (Baesens 2014). Our discovery-driven analytics for brand advertising based on large-scale social data makes the following contributions in the area of big data.

- The key contribution of this paper is a novel framework for social brand advertising using implicit brand–brand networks. We define undirected and directed weighted networks to capture brand relationships. Network information is then used together with sentiment analysis to identify target audiences. To leverage multiple types of information from the big data that forms the basis for this

[3] http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-041014.

work, the framework involves multiple components, which include a hierarchical community detection algorithm to identify sets of closely related brands, and an algorithm for selecting important/influential brands from the set of closely related brands.

- To help investigate network properties, we propose a network normalization technique to capture both local and global information. In addition, we define new structural measures for analyzing these networks by modifying traditional network measures such as degree and eigenvector centrality to incorporate weights.

- Our study uses a very large (approximately 2.1 TB) real-world dataset to test and demonstrate the effectiveness of our methods. To obtain high-quality data, we design effective rules to filter out spam users and their corresponding activities.

- To allow processing of large-scale data and networks, we develop and implement several MapReduce-based algorithms for network generation, processing, and brand influence ranking.

- Finally, we design a novel evaluation technique to test the effectiveness of our targeting framework. To identify target audiences who are not yet engaged with a focal brand, we split the data into two time periods, with the first used for audience selection and the latter used for testing. This is used to check if a user identified as a potential target for a focal brand based on data in the selection period is engaged with this brand in the testing period. We evaluate performance improvements from different components of our targeting framework.

The rest of this paper is organized as follows. First, we review related work. We then introduce the overall framework and explain the data and data cleansing. Next, we introduce implicit brand–brand networks, describe how they are generated and normalized, and lay out important network measures used later for analyzing the networks. This is followed by a description of our proposed targeting framework and the accompanying algorithms for target audience discovery. We explain experimental results and evaluation, followed by conclusions and directions for future work.

## Related Work

With rapid developments in online social media, there is growing interest around the influence of networks on behavior, and how behaviors spread across connected individuals (Centola 2010; Gruzd and Wellman 2014). While some studies note the importance of strong ties over weak ties for generating influence in social networks like Facebook (Bakshy et al. 2012; Bond et al. 2012), researchers find that repeated exposure to social information through weak ties is also effective for social influence; repeated exposure to interests and preferences of weak network ties are noted to be common in social network platforms (Kwon et al. 2014). With growing activity around social media and consumer networks, there is a great deal of interest among marketing managers and researchers in leveraging network-based information. Multiple studies have noted the value of social influence for viral marketing (Bruyn and Lilien 2008; Domingos and Richardson 2001) and pushing product adoption ( Godes and Mayzlin 2009; Manchanda et al. 2008). Research in these areas draws on developments in different disciplines, and Hill et al. (2006) provide a good survey.

Traditional target marketing is largely based on demographic, behavioral, and attitudinal data, and researchers find that information on social ties also has predictive value (Goel and Goldstein 2013). Research has examined social targeting or behavioral targeting which refers to learning from past user behavior, especially user feedback (such as comments and clicks), to find the best match between users and advertisements. Research on social advertising targeted at friendship-networks of "fans" on a Facebook page for a charity finds these to be more effective than un-targeted or demographically targeted ads (Goldfarb and Tucker 2011). In another Facebook-based study, Lee et al. (2014) analyze large-scale data to examine the design of effective social advertising content. In the area of audience selection, Provost et al. (2009) show that user profiles can be built from co-visitation patterns of social network pages. Brand affinity is inferred from users' observed common actions on brand content. They also suggest a holdout testing based evaluation framework for brand advertising, somewhat similar to that used in our paper.

Social networks are being used for promoting brands and developing brand communities (Fournier and Lee 2009). Research has also examined online consumer interactions (De Valck et al. 2009) and the use of social networks to foster consumer engagement (Brodie et al. 2013). Social network-based communities are now seen as central to how brands interact with consumers, with brand fans noted to be more emotionally connected to the brand and prone to greater positive word-of-mouth (De Vries et al. 2012).

Consumers can influence each other in many ways via online social networks. These include online product reviews and

feedback, communication of purchase/adoption decisions, product recommendations, and interactions in social brand communities (Mangold and Faulds 2009). Constructing and analyzing implicit networks from such interactions have recently attracted research attention. Zhang et al. (2014) build large implicit brand–brand networks from Facebook fan pages to investigate whether influential brands have a large number of fans and receive more positive comments from social users. Chau and Xu (2012) analyze blog content to identify implicit networks from blog subscriptions and comments. They use network analyses and clustering to gather business intelligence from blogs.

Text mining is useful for automated analyses of user-generated content (Aral and Walker 2011). Identifying social sentiment by incorporating user-level information in social networks can improve accuracy (Tan et al. 2011). There has been a wide range of research done on sentiment analysis, from rule-based, bag-of-words approaches to machine learning techniques (Pang et al. 2002). In this paper, we deploy a sentiment engine implemented in previous research to identify user positivity based on user's comments (Zhang et al. 2011). We thus propose a framework that combines the use of network analysis and sentiment analysis to identify potential targets for brand advertising.

## Overall Framework

Our proposed framework for identifying the target audience has four phases. The first phase is to construct two weighted brand–brand networks based on user historical activities on a social media platform. These networks are then normalized from both global and local perspectives for the further analysis. The second phase is to identify a set of brands that are closely related to a focal brand. For this, we have developed a hierarchical community detection algorithm that works on the undirected and weighted brand–brand network built in the first phase. The third phase is to obtain a subset of important brands from the set of brands in the second phase. For this, we have developed a brand importance identification algorithm based on the directed and weighted brand–brand network built in the first phase. The fourth phase selects target users (audiences) from the set of important brands closely related to the focal brand, based on sentiment of users' comments made across all brands. Finally, we design an evaluation technique to test the effectiveness of our proposed targeting framework. Figure 1 shows the overall framework, which we now describe in detail in the following sections.

## Dataset

We collected a large (approximately 2.1 TB) dataset from Facebook. In this section, we describe the details of the data collection, preprocessing, and cleaning performed to generate a clean dataset for analysis.

### Data Collection

Facebook, the largest and most popular social network platform, has more than one billion accounts. Many companies, organizations, communities, and individuals build their own pages on Facebook to share and communicate with their fans. Facebook is also a popular social marketing website for advertisers and their marketing partners to reach users based on their demographics, activities, and interests. The extensive amount of textual and interaction information generated by users has made it a promising platform for social online brand advertising. In this work, our focus is on top social brands as our object of analysis (i.e., the brands with a large number of user activities). We use Facebook Graph API[4] to download all activities visible on a brand page such as posts by the brand administrator, as well as posts by users, such as comments, likes on posts, and some public user profile information (only gender and locale). Note that the "share" button was launched in late 2011, hence we do not use it in this paper because of lack of data consistency over the entire time period of analysis. Also, detailed user profile information such as demographics is not available from Facebook APIs. Each brand may have various numbers of posts depending on its posting frequency. These posts may be text, images, videos, links, or a combination of these. A post is any information that the brand wants to share and interact with users. For instance, posts may be about a new product release, company annual report filing announcement, special day greetings, surveys, or other important events and activities. Any Facebook user can respond to these posts by liking or making comments on them. While there is no "dislike" functionality on Facebook, textual responses to posts can be used to indicate positive as well as negative opinions. The dataset (shown in Table 1) used for this study was collected from January 1, 2009, through January 1, 2013. It contains data from 13,808 brand pages and approximately 280 million users. It covers data from brands in 122 countries in 172 categories as described by Facebook's classification system. For example, brand "United" (United Airlines) from the "Travel/Leisure" category has 174,881 unique users who have interacted with the page as of the given date, as well as 1,461 posts and 151,568 comments.

[4]Facebook Graph API (https://developers.facebook.com/docs/graph-api/).

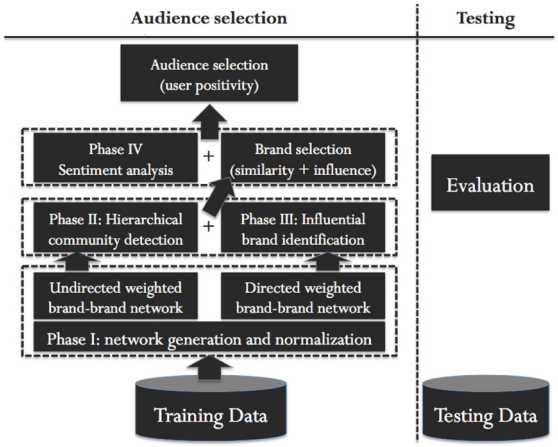
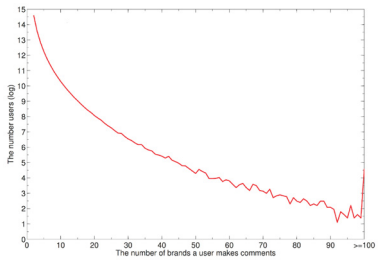


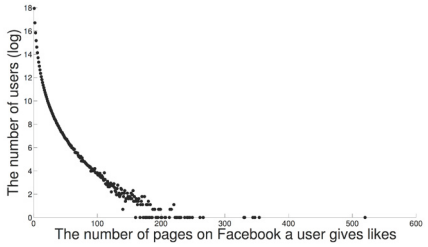

**Figure 1. A Framework for Audience Identification and Evaluation**

**Table 1. Dataset Description and Statistics**

| | |
|---|---|
| Number of brands | 13,808 |
| Number of unique users | 286,862,823 |
| Number of brand regions (countries) | 122 |
| Number of brand categories | 172 |




**(a) User Comments Distribution**




**(b) User Likes Distribution**

**Figure 2. Distribution of User Comments and Likes**

### *Data Cleansing*

Data quality is of paramount importance in any analytics study as it can affect the performance and results. First we removed brands in which most of the posts and comments were not in English, because sentiment identification used in our framework for non-English text is not well understood and accuracy is not guaranteed. To produce robust results we ignored users who made very few comments (i.e., less than five) across all brands, as user opinions drive the measurement of user positivity, as explained later in the "User Sentiment Identification" subsection. We then designed a set of rules to remove fake users and their corresponding activities. Our data shows that, on average, a user comments on four to five pages and likes posts on seven to eight pages as shown in Figure 2. They exhibit scale-free distributions. In Figure 2(a), we aggregate all Y-axis values for all values greater than 100 on the X-axis. Users connecting to an extremely large number of brands are likely to be spam users or bots. For example, we found one spam user who appeared on more than 600 different brands. We also detected one user who "liked" posts across 520 different brands. As most users are likely to be interested in a small number of brands, we discarded users making comments on more than 100 brands

| Table 2. Cleaned Dataset of Top 2,000 Brands | | |
|---|---|---|
| After Cleaning | | After Selecting Top Brands |
| Number of brands | 7,580 | 2,000 |
| Number of unique users | 97,699,832 | 16,306,977 |
| Number of comments | 2,327,635,302 | 470,742,158 |
| Number of positive comments | 651,231,870 | 179,009,470 |
| Number of negative comments | 234,571,177 | 60,613,968 |
| Number of brand categories | 150 | 118 |
| Number of posts | 13,206,402 | 3,793,941 |

and those liking posts on more than 150 brands. In addition, we detected other kinds of spam users. For example, there was one user who liked 7,963 posts out of all 8,549 posts for a brand. We assume that it is likely to be a spam user if this ratio is very high. On average, our data shows that a user likes 0.105% posts of a brand page. Therefore, we set this ratio threshold to be 90% for every user except the page administrator. Finally, we also removed users who posted many duplicate comments containing URL links, which sometimes direct to phishing sites. For instance, a test on Barack Obama's page found 209,864 duplicate comments out of 2,987,505 in total. The dataset for our analyses is from the top 2,000 brands, where "top" refers to brands having the largest numbers of user activities on their Facebook pages. The statistics of our cleaned dataset are shown in Table 2.

## Network Analysis

From the cleaned dataset, we built two different implicit brand– brand networks based on user activities. The first is an undirected and weighted network, and the second, a directed and weighted network. Edge weights in both networks were normalized. These two networks are very large, and hence we implemented a MapReduce-based algorithm (described in Appendix A) to construct them. We then modified the standard structural property measures for networks and used them to analyze our brand–brand networks. These structural properties include (weighted) degree and eigenvector centrality measures, which consequently lead us to our targeting framework.

### *Undirected and Weighted Brand–Brand Network*

Each brand has various properties such as its category as defined by Facebook, number of fans (fan indicates a user who is engaged with a brand page irrespective of whether they express positive or negative sentiments toward that brand.), number of people "talking about" it, and a record of users' activities. The activity information is used to capture the implicit relationships among brands and extract the brand–brand network. In this network, brands are designated as nodes, and an edge between two nodes is created if there are common users with activity on both brands (i.e., users who comment on or like posts made by both brands). The larger the number of common users across two brands, the higher is the weight of the edge between these brands. This network represents the brand affinity. We define an undirected and weighted brand–brand network (denoted as $B$) as $B = <V, E>$, where the set of nodes $V$ correspond to brands and the set of edges $E$ carry weights that represent the number of common users between any two nodes. Formally, $V = \{b_i\}$ with $b_i$ being a brand having $F_i$ as the set of users who have activities on this brand, and $E = \{(b_i, b_j) \mid F_i \cap F_j \in \Phi\}$ with corresponding weight $w_{ij} = |F_i \cap F_j|$. Here, $1 \le i, j \le N$, and $N$ is the total number of brands, which is 2,000 in this study. Alternatively, for the convenience of explaining network measures in the following section, we use the adjacency matrix $A$ to represent the network $B$ as

$$A_{ij} = \begin{cases} w_{ij} & \text{if node } j \text{ connects to node } i \\ 0 & \text{otherwise} \end{cases}$$

where $w_{ij}$ is the weight of the edge between brands $b_i$ and $b_j$, representing the number of common users with activities on both brands $b_i$ and $b_j$.

**Normalization of brand–brand network: ($B \rightarrow B$)**: Well-known and more popular brands typically attract more users and have larger amounts of user activity. Such brands with larger numbers of active users then have higher numbers of common users with connected brands, compared with brands that not as well-known. This leads to the more popular brands having much larger edge weights in the brand–brand network. These higher weighted edges associated with a few very popular brands can then dominate analyses in the network.

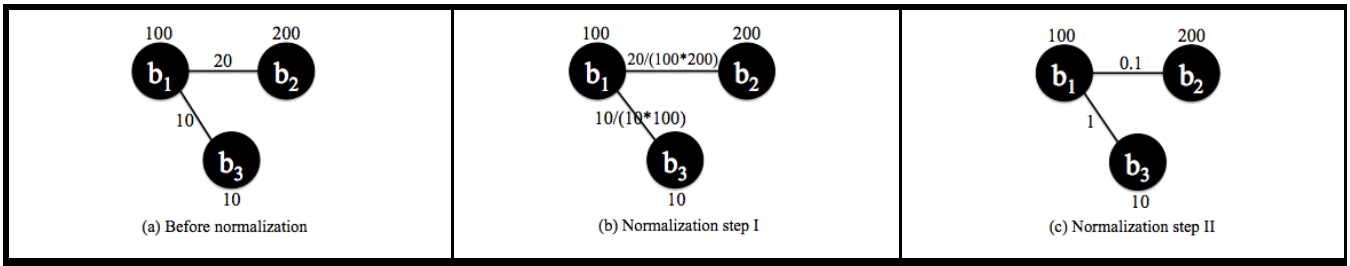



**Figure 3. Undirected Network Weights Normalization**

To facilitate comparison across brands in the network, the edge weights need to be normalized. A simple approach is to normalize with respect to the global maximum weight in the network. However, we then lose global network semantics such as the distribution of connection strength among edges of a brand relative to the size of a brand. Consider the case shown in Figure 3(a). Brand $b_1$ has 100 active users. Brand $b_2$ has 200 active users, and brand $b_3$ has 10 active users. The number of common users between $b_1$ and $b_2$ is 20. The number of common users between $b_1$ and $b_3$ is 10. If we normalize globally, the connection ($b_1$, $b_2$) will appear stronger than the connection ($b_1$, $b_3$), since the weight $w_{12}$ is greater than the weight $w_{13}$. However, the connection ($b_1$, $b_3$) is actually relatively stronger than the connection ($b_1$, $b_2$), because all (100%) of the brand $b_3$ users are connected to $b_1$, while only 10% of $b_2$ users are also interested in $b_1$. Therefore, we propose a different normalization technique to characterize the strength of an edge in the normalized network (denoted as $B_n$). See the example in Figures 3(b) and 3(c). The normalization for network $B \rightarrow B_n$ occurs in two steps.

**Step I.** We normalize an edge weight between two brands $b_i$, $b_j$ by setting the weight

$$w'_{ij} = \frac{w_{ij}}{f_i * f_j}$$

where $f_i$ and $f_j$ here are the number of active users for brands $b_i$ and $b_j$, respectively. $w_{ij}$ is the un-normalized weight of the edge between $b_i$ and $b_j$ (number of common users who have commented on posts made by both brands $b_i$ and $b_j$).

**Step II.** We then normalize all $w'_{ij}$ by setting

$$w^*_{ij} = \frac{w'_{ij}}{\max_{\forall(i,j)}\{w'_{ij}\}}$$

## Directed and Weighted Brand–Brand Network

The normalized network $B_n$ allows a global comparison of relationship strength among brands. However $B_n$ has two problems: (1) $B_n$ does not indicate the relative affinity of brand $b_i$ to its neighbors, and (2) for an edge between two brands $b_i$ and $b_j$, $B_n$ does not distinguish the affinity from the perspective of each brand, that is, the affinity of $b_i$ to $b_j$ from $b_i$'s perspective and the affinity of $b_j$ to $b_i$ from $b_j$'s perspective (they are taken as the same in the undirected networks $B$ and $B_n$). For example, considering the un-normalized network in Figure 3(a), we see that 20% and 10% of $b_1$ users have activities on $b_2$ and $b_3$ respectively, while all $b_3$ users have activities on $b_1$. Hence from the perspective of $b_1$, $b_3$ is closer to it than $b_2$, even though $b_1$ has more common users with $b_2$ than with $b_3$. To address this, we define a directed and weighted brand–brand network (denoted as $\vec{B}$). $\vec{B}$ is almost the same as $B$ except that edges become directed. The weights of the edges in both directions are equal before normalization (i.e., $w_{ij} = w_{ji}$ = the number of common active users between brand $b_i$ and brand $b_j$).

**Normalization of brand–brand network**: ($\vec{B} \rightarrow \vec{B}_n$): The edge strengths between nodes in the directed network are normalized in a way that captures local information about a brand and its immediate neighbors. Figure 4 shows the process of generating the normalized directed network (denoted as $\vec{B}_n$). The normalization for $\vec{B} \rightarrow \vec{B}_n$ is performed as follows: For each individual edge between two brands, $b_i \rightarrow b_j$, we update the weight using

$$w'_{ij} = \frac{w_{ij}}{f_i}$$

where $f_i$ here is the number of active users for brand $b_i$.

## Network Generation

Having defined all the networks (i.e., $B$, $B_n$, $\vec{B}$, and $\vec{B}_n$), we now focus on the process used to generate the network containing common users between brands. The raw data downloaded and aggregated from Facebook consists of triplets: <$brand_{id}$, $user_{id}$, # of activities>. The size of the file is too

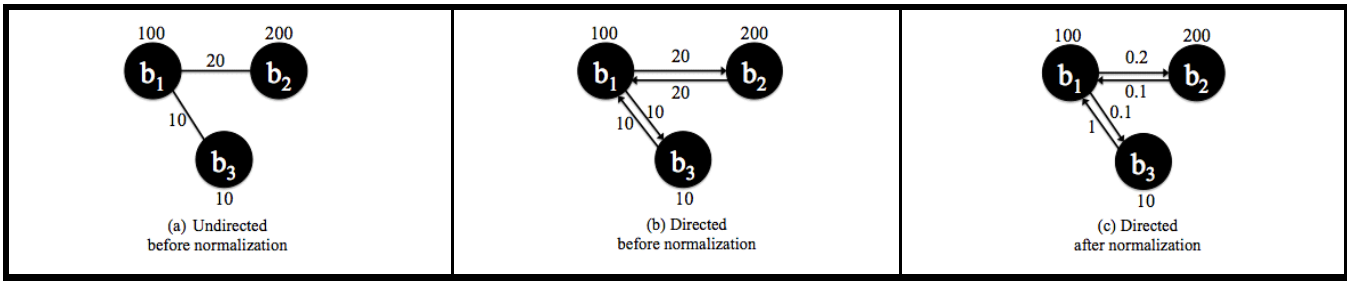
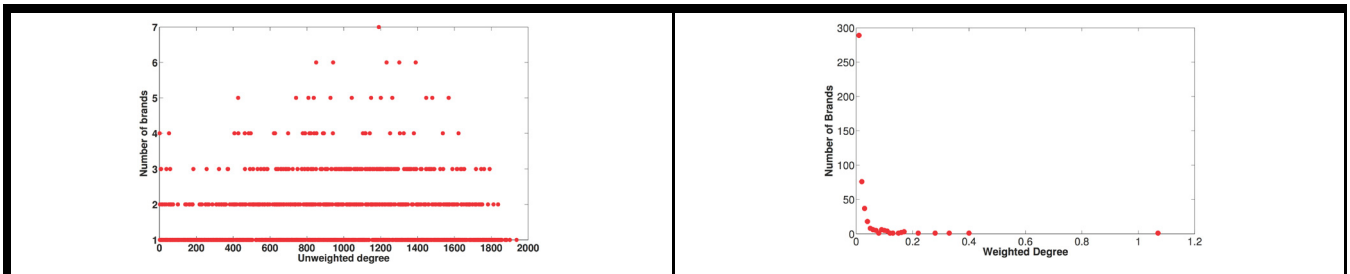



**Figure 4. Directed Network Weights Normalization**

**Table 3. Properties of Normalized Networks**

| Property | Undirected Network $B$ | Directed Network $\vec{B}_n$ |
|---|---|---|
| Number of nodes | 2,000 | 2,000 |
| Number of links | 965,605 | 1,931,210 |
| Average weighted degree | 0.662 | 1.767 |




**(a) Unweighted** **(b) Weighted**

**Figure 5. Network Degree Distributions**

large to be processed by a single machine. For example, to get common users between two brands $b_i$ and $b_j$, we need to do intersections between two sets $S_i$: {all users in $b_i$} and $S_j$: {all users in $b_j$}. This consumes enormous processing time because each brand typically has millions of unique users who have activities on its page. We used Hadoop to efficiently generate our network file in the format of <$b_i$, $b_j$, # of common users>. The basic map and reduce functions are shown in Algorithm 1 (see Appendix A). Without using Hadoop and other distributed computing techniques, it would have been impossible to even load such a large dataset (approximately 2.1 TB) into one single machine.

## Network Measures

There are many structural properties for networks, including node degree, network diameter, density, clustering coefficient, and centrality. In this paper, we focus on two important properties, node degree and eigenvector centrality, which drive our targeting approach. Most of these structural network measures have been defined and studied on unweighted networks. In this work, we extend these metrics for weighted networks.

**Node Degree**: The simplest yet most frequently used property of a node is its degree (i.e., the number of connections it has to other nodes). The degree of node $i$ ($b_i$) can be easily computed from the adjacency matrix $A$: $k_i = \Sigma_j A_{ji}$. $A$ can represent a weighted network ($A_{ji}$ could be any positive integer) or an unweighted network ($A_{ji}$ is either 1 or 0). Table 3 shows the basic properties of network $B_n$ and network $\vec{B}_n$. Larger connection strength means higher brand affinity. Figure 5(a) shows that some brands/nodes have a small number of neighbors while others have a large number of neighbors. The average weighted degree of 0.662 is the average connection strength of neighbors. Figure 5(b) shows the degree distribution for the weighted network indicating a

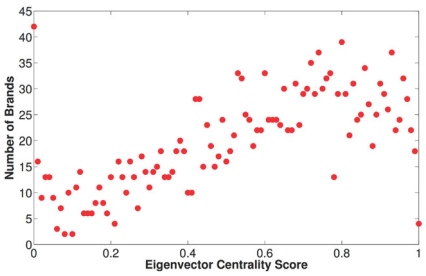



**Figure 6. Brand–Brand Network Eigenvector Centrality Distribution**

scale-free network. This implies that there are a few brands with very high weighted degree and lots of brands with low weighted degree.

**Eigenvector Centrality**: It is used to measure the influence/importance of a node in a network. It is based on topological features alone and takes into account only information about the neighborhood of a node. It assigns relative scores to all nodes in the network based on the idea that connections to more important nodes contribute more to the importance of the node in question than connections to less important nodes. Since our brand–brand network $B_n$ is weighted, we modify the original eigenvector centrality measure as follows. For the given network $B_n$, let $A = (w_{ij})$ be the adjacency matrix. By incorporating edge weights, the eigenvector centrality score $c_i$ of each brand $b_i$ can be defined as

$$c_i = \frac{1}{\lambda}\sum_{j\in N(i)} w_{ij}c_j = \frac{1}{\lambda}\sum_{j\in B_n} w_{ij}c_j$$

where $N(i)$ is a set of the neighbors for brand $b_i$ and $\lambda$ is a constant. This calculation can be rewritten in vector notation with a small mathematical rearrangement as the eigenvector equation: $Ac = \lambda c$.

The eigenvector centrality distribution for our brand–brand network $B_n$ is shown in Figure 6. It reveals that brands in the network have influence scores distributed widely between [0, 1]. There are 42 isolated brands and about 15 brands with influence scores close to 0 in the network. The rest of the brands have eigenvector centrality scores between 0 and 1, meaning that they can have either strong connections or weak connections to other brands. This centrality measure is useful for ranking brands with respect to the focal brand and helps identify important brands that can attract larger social audiences. However the regularly used eigenvector centrality measure does not take node weights, link weights, and directionality of edges into account. We propose a brand importance-ranking algorithm to incorporate these weights in the "Brand Ranking" subsection.

## Proposed Framework for Audience Selection

In this section, we describe our overall framework for identifying target audiences for online advertising. The framework consists of four major phases: the first phase is to construct and normalize weighted brand–brand networks derived from user activities. The second phase is to find a set of closely related brands to the focal brand and the third phase is to calculate global influence of brands and rank them. Then we combine the second and the third phases to obtain a set of closely related influential brands. From these brands, the fourth phase is to select users for targeting (audiences). For the second phase, we propose a hierarchical community detection algorithm for finding a set of closely related brands. For the third phase, we propose an algorithm for ranking brands to select the most important/influential brands. These two algorithms are integrated for the brand selection. In the fourth phase, we identify users for targeting, based on sentiments of comments made across all brands.

### *Hierarchical Community Detection*

Most users on social media platforms evolve in their level of activity on various brands. For instance they start following pages, post comments periodically, and/or like posts on brands in which they are interested. Moreover, their activity level changes over time. Our framework allows us to find the audience who may be potentially interested in the focal brand from the population of users ($u$) who are not engaged with a focal brand ($F$) but are engaged with other brands ($b$). We assume that if $b$ is closely connected to $F$, then $u$ will be interested in $F$ with a high probability. Thus our problem first involves finding brands closely related to the focal brand, which can be addressed using community detection algorithms.

The community detection algorithm is typically formulated as finding a partition $C = \{C_1, C_2, \ldots, C_k\}$ of a network $N = (V, E)$,

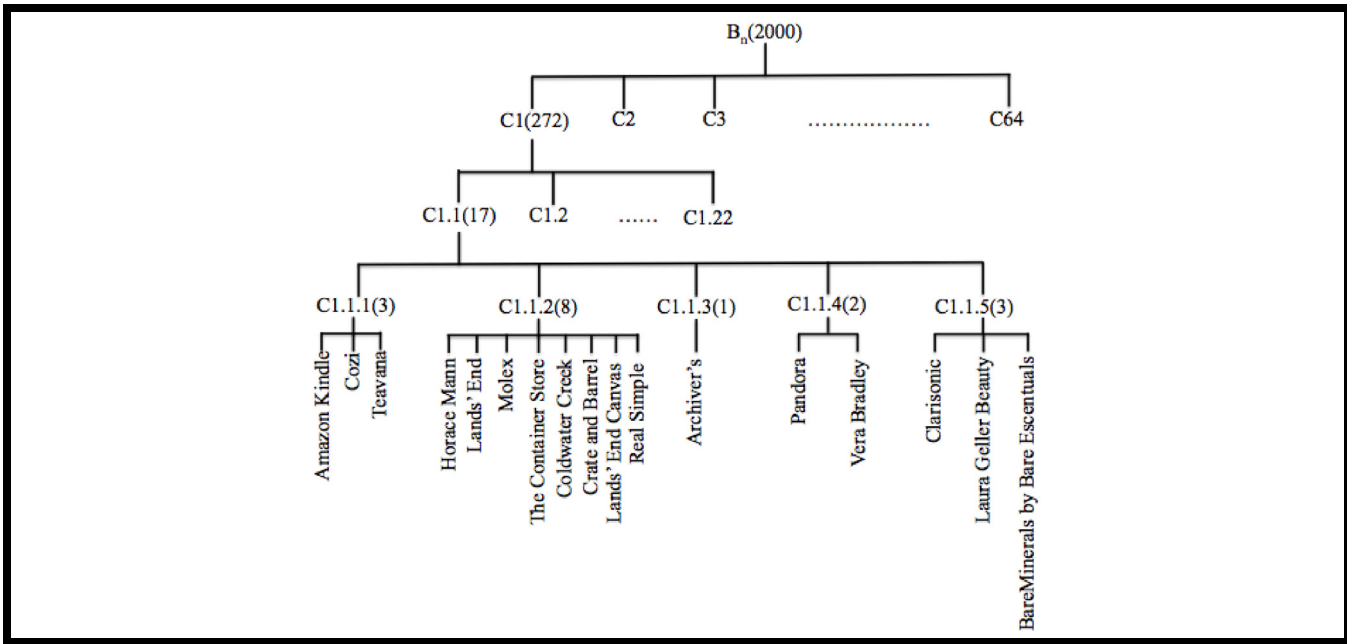



**Figure 7. Hierarchical Community Detection for Undirected Brand–Brand Network**

where $i, j, \forall i, j, C_i \cap C_j = \varnothing$. Here, $k$ is the number of communities. Our approach in this work is based on a well-known community extraction algorithm called modularity maximization (Newman 2006). Modularity ($Q$) is a standard quantitative measure of the quality of a partition in a network. Higher modularity values indicate communities having higher numbers of intra-community links as compared to inter-community links. The modularity maximization algorithm is used to find the community assignment for each node such that the modularity is maximized. According to Newman (2006), maximum modularity does not mean that a network necessarily has a community structure. In particular, this is true if the communities are cliques. Therefore, using modularity to extract communities may result in large communities, which in turn could be comprised of smaller communities. Figure 7 shows that the initial 64 partitions $C_1, C_2, \ldots, C_{64}$ are generated when the community detection algorithm runs once on the undirected and weighted brand–brand network.

Since these communities are very large, we present a new approach that can further divide these bigger communities into smaller, focused communities. While extracting small communities from each of the bigger communities, we consider only the subnetwork that contains those nodes that belong to the bigger community. The reason is that the links connecting a bigger community to other communities have already been considered while identifying the previous communities. Our algorithm is recursive. At the beginning of each round, we run the modularity-based algorithm to generate communities. Then the algorithm continues to subdivide big communities or declare it as a final community, depending on the stopping criteria. It stops dividing a community when the size reaches a predefined threshold. Within each dividing round, all big communities can be divided in parallel. The algorithm is described in Appendix B (Algorithm 2). The threshold we used in this paper is 10, based on experimentation to obtain tightly knit communities.

We applied the community detection algorithm to the undirected weighted brand–brand network resulting in 64 communities initially. Figure 7 shows the hierarchical community detection process for the normalized undirected and weighted brand–brand network $B_n$. The number in parentheses represents the number of brands within that subcommunity. For example, "Amazon Kindle" is a member of a community along with 271 other brands. This subcommunity contains 13.6% brands of the entire network. A new network is constructed based on these 272 brands and reclassified into 22 communities. The size of the community containing "Amazon Kindle" reduced to 17 at this point. The last community detection iteration splits 17 brands into 5 different communities as shown in Figure 7. Figure 8 shows the largest community in the network $B_n$. It has 272 brands. The size of a node represents the node degree and the color of a node represents the membership of subcommunity. Bigger indicates higher degree. Figure 9 shows one of the subcommunities, related to airlines and hotels, extracted from the largest community in Figure 7.

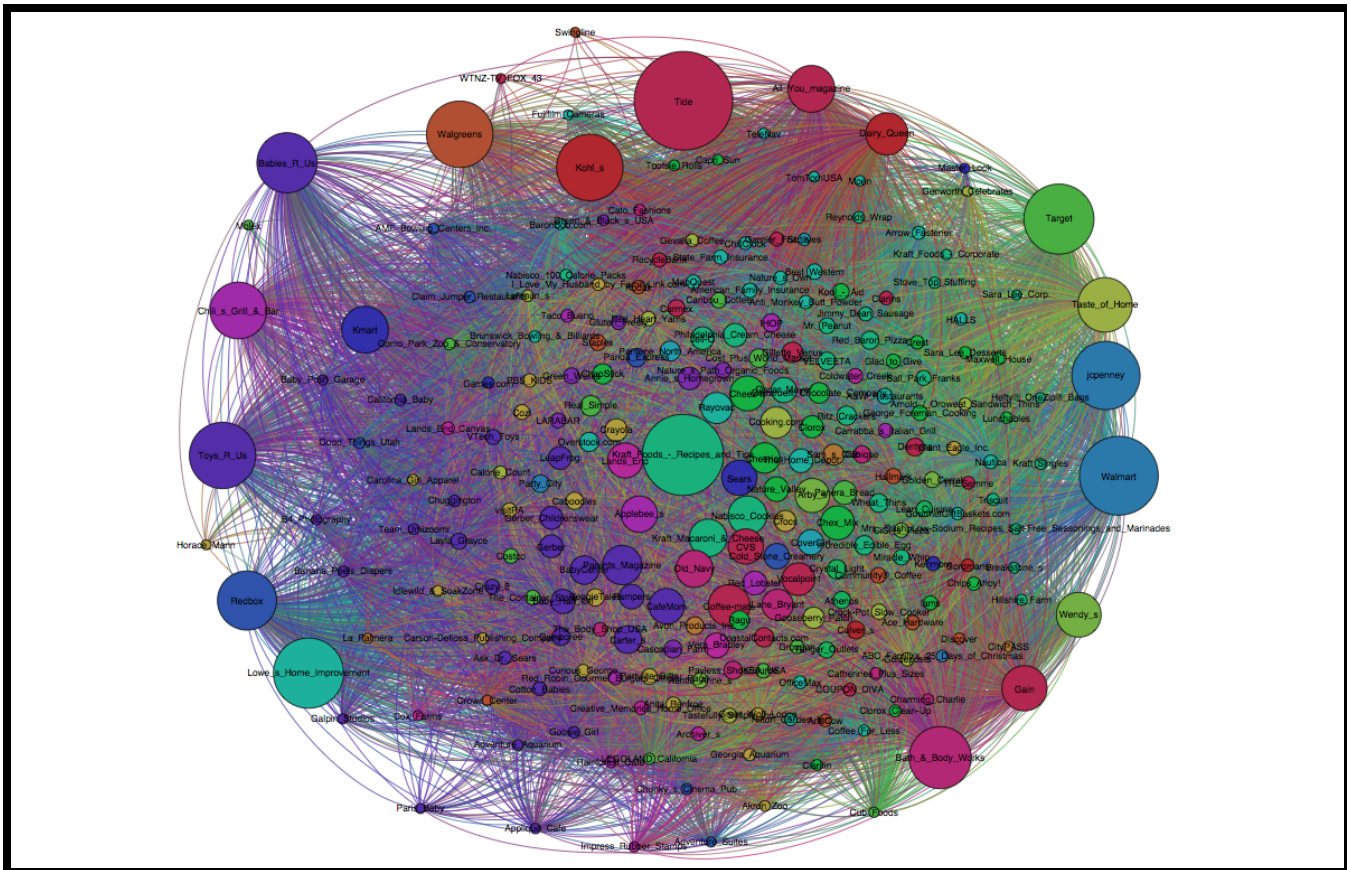



**Figure 8. The Largest Community in the Brand–Brand Network**

**Figure 9. A Sample Subcommunity**

### Brand Ranking

Once we obtain a set of brands, which are in the same community as the focal brand, we need to choose the most important brands based on some ranking criteria. We propose a brand importance-ranking algorithm (denoted as *bRank*) based on the concept of PageRank using the eigenvector centrality computation discussed in the previous section. For this step we use the normalized directed and weighted network $\vec{B}_n$. The major difference from PageRank is that we incorporate both node weights and edge weights. The basic idea of *bRank* is as follows: (1) The brand with a large number of user activities is considered to have higher importance. (2) The brand connected to an important brand will contribute more to its own importance score than one that is connected to a less important brand. The ranking algorithm on the network $\vec{B}_n$ can be formally defined as follows:

$$bRank(b) = \left[ (1-\beta) + \beta * \sum_{i=1}^{i} 1\{b_i, b\} * bRank(b_i) * C_e(b_i) \right] * C_n(b)$$

where

- $\beta$ is the damping factor, the probability of following a link at random. It is used to solve the problem of dead ends (i.e., brands without out-links). As suggested in the literature, the common range is [0.8, 0.9]. In our experiment, we choose 0.85.
- $bRank(b)$ is the brand ranking of brand $b$.
- $bRank(b_i)$ is the brand ranking of brand $b_i$ and $l$ is the number of incoming links on brand $b$.
- $\mathbf{1}\{b_i, b\}$ is an indicator function. It is 1 if there is a link from $b_i$ to $b$, 0 otherwise.
- $C_n(b_i) = \dfrac{W_v(b_i, b)}{\sum_{j=1}^{m} W_v(b_i, b_j)}$, where $m$ is the number of outbound links on brand $b_i$, $b_j$ are the brands pointed to from $b_i$ and $W_e(b_i, b_j)$ is the weight of the edge $(b_i, b_j)$. It is the edge weight contributor to the ranking of brand $b$.
- $C_n(b) = \dfrac{W_v(b, b)}{\sum_{t=1}^{n} W_v(b_t, b_t)}$. It is the node weight contributor to the ranking of brand $b$. $W_v(b_t, b_t)$ is the size of brand $b_t$.

This *bRank* algorithm is essentially similar to solving the eigenvector of a matrix, which takes considerable time if the matrix dimension is high. To improve performance, we devised and implemented a distributed algorithm using MapReduce (Algorithm 3 in Appendix C). The initial importance score for each brand is set to be 0.5. We iteratively run the algorithm until convergence. For brands in the same subcommunity as the focal brand "Amazon Kindle" in our brand–brand network (2,000 brands), the decreasing order of their importance scores is *Cozi* > *Teavana* > *Horace Mann* > *Lands' End* > *Molex* > *The Container Store* > *Coldwater Creek* > *Crate and Barrel* > *Lands' End Cavas* > *Real Simple* > *Archiver's* > *Pandora* > *Vera Bradley* > *Clarisonic* > *Laura Geller Beauty* > *BareMinerals by Bare Escentuals*.

### User Sentiment Identification

Users tend to express their opinions positively, neutrally, or negatively through their comments. Many easy-going users tend to make nonnegative comments or like other's posts, while some tough users like to leave nonpositive comments. The purpose of brand advertising is to expand their marketing influence and increase customers. Positive people are more likely to spread more positive information among their friends than people with a high degree of negativity. Prior research has developed effective ways to identify user positivity on social media platforms, using ensemble learning methods to mine characteristics of natural language and social media texts (Zhang et al. 2011). Here, we use this sentiment analysis method that has been demonstrated to be suitable for user comments in social media. It was tested on Facebook comments and Twitter tweets and seen to achieve an accuracy of 86%. We use a simple way to identify users' positivity by calculating the positive ratio of all historical comments made by a user. Since we already removed users who made only a very few comments during the data cleaning process, this approach for audience selection works well. Positivity of a user $u$ (denoted as $POS_u$) is defined as

$$\mathrm{POS_u} = \frac{\text{Number of positive comments}}{\text{Number of positive comments + Number of negative comments}}$$

We ignore neutral comments here because they do not express any opinions, but appear to just state facts.

### Audience Selection

The goal in online advertising is to find potential users who can be targeted for ads. Thus far, we have developed techniques to find closely related brands, calculate brand importance scores, and identify user positivity. In this section, we integrate these three steps to select target audiences. Given a

focal brand $f$, an undirected and weighted normalized network $B_n$, and a directed and weighted normalized network $\vec{B}_n$,

- We first perform hierarchical community detection on $B_n$ to find closely related brands $\{b_i \in C^*\}$ from the community $C^*$ that includes the focal brand $b_f$.
- We next use the *bRank* algorithm to calculate influence scores of all brands in the network $\vec{B}_n$. We then rank all brands $b_i \in C^*$ and obtain the top $k$ influential brands as our final selected brands: $b_1, b_2, \ldots, b_k$.
- We consider all users from these $k$ brands as a candidate audience pool for the focal brand $b_f$:

$$U_{b_f} = \bigcup_{i=1}^{k} U_{b_i'}$$

  where $U_{b_j}$ are the users from brand $b_j'$.

- Finally, we rank each user $u_j \in U_{b_j}$ based on user positivity $POS_{u_j}$ defined by the positive ratio of historical comments he/she made across all brands. We can choose top $m$ users as the final target audiences. Here, $m$ and $k$ are two control variables for selecting a fixed number of target users. We can also adjust the size of the community $C^*$ if necessary through how deep the hierarchical community detection goes.

## Analysis and Results

In this section, we first present our evaluation techniques used to test the effectiveness of our targeting framework as compared to some specific baselines. We evaluate performance on a sample of focal brands. All experimental data collection and cleaning were based on the dataset discussed in the "Dataset" section. The sequential algorithms were executed on a single-node machine. The distributed algorithms were executed on a Hadoop cluster with 10 nodes with each node having 4GB memory.

### Empirical Results

Before discussing the effectiveness of our targeting framework, we first present empirical results from phases III and IV. Results from phase II (i.e., hierarchical community detection) have already been discussed in the previous section. Table 4 shows the 10 most influential and 10 least influential brands along with their categories. A rank of "1" indicates the most influential brand and a rank of "-10" indicates the least influential brand. These are derived from our ranking algorithm on the network $\vec{B}_n$ developed in phase III (i.e., brand importance ranking). The individual brand ranking score is obtained by applying a MapReduce-based brand-ranking algorithm on a large normalized directed and weighted brand–brand network (2,000 brands). It was executed for 1,000 iterations on a Hadoop cluster. We compared the execution time with a sequential algorithm on a single-node machine. These execution times were approximately 153 seconds (less than 3 minutes) for the MapReduce implementation as compared to 1,127 seconds (almost 19 minutes), respectively. This demonstrates the capability of our MapReduce-based algorithm to perform efficiently on large networks.

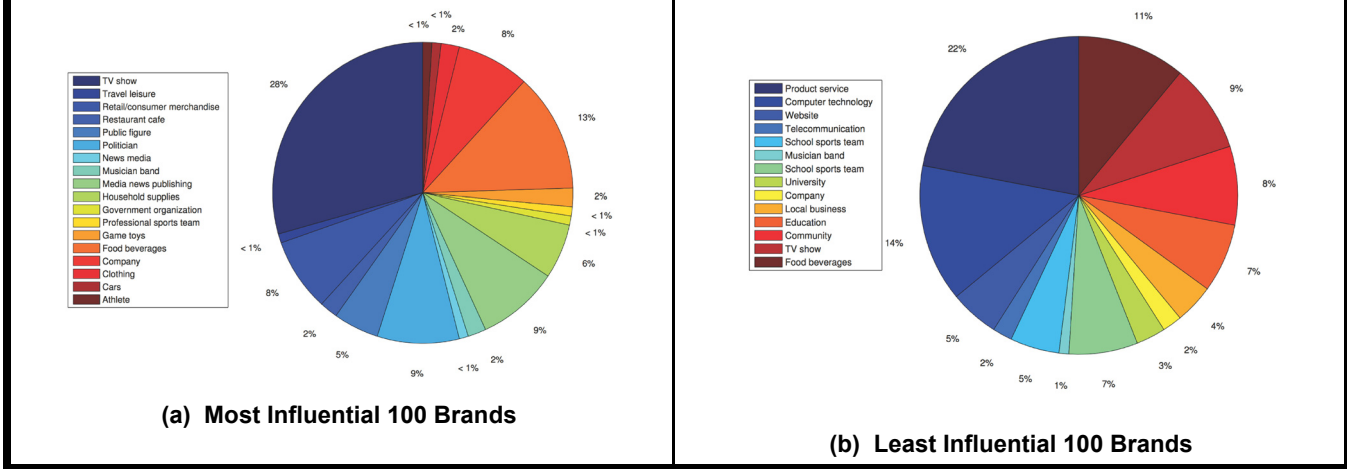
Each brand has a category defined by Facebook, such as, sports, politician, food beverages, clothing, or TV show. We found that among the top 100 most influential brands, 28 brands are TV shows, 13 are food/beverages, and 8 are musician bands. Similarly, among the 100 least influential brands, 22 brands are product service, 14 are computer technology, and 11 food/beverages (see Figure 10).

### Evaluation Process and Results

The evaluation process is to test the effectiveness of our model to find potentially interested audiences for a focal brand. Obviously, the best evaluation technique would be to launch advertisements on a social platform (e.g., Facebook) based on our proposed framework and some other baselines for the focal brand and then compare click-through rates and conversion rates for both during a specific time period. A "live" experimental study on this will be the subject of a separate future investigation. In this paper, we use a different approach for evaluating the audience selection framework. The key objective of evaluation is to examine whether the audience we selected is really interested in the focal brand or not. Most social platforms, like Facebook, have mechanisms to measure if a user is interested in a brand or not, based on one or more of the following actions: (1) becoming a fan of that brand, (2) liking campaigns posted by that brand, and (3) making positive comments on that brand. Due to limitations of the Facebook Graph API, we cannot get individual fan profile information for each brand. We can only extract the total number of fans. We do not have data on when a specific user becomes a fan of the brand, but we do have time stamps for their likes and comments on posts. Hence we use information on user actions (2) and (3) for our evaluation. The general evaluation framework is shown in Figure 11. We have the entire dataset from time $0$ ($T_0$) to time $t$ ($T_t$). We split this dataset into two stages in time order: selection $<T_0, T_i>$

**Table 4. Top and Bottom Ranked Brands**

| Rank | Brands | Category | Rank | Brands | Category |
|---|---|---|---|---|---|
| 1 | Barack Obama | Politician | -1 | Molex | Computer technology |
| 2 | NPR | Media news publishing | -2 | Google+ | App page |
| 3 | CNN | Media news publishing | -3 | Dentyne | Product service |
| 4 | Starbucks | Food beverages | -4 | NAVIGON | Product service |
| 5 | Justin Bieber | Musician band | -5 | Vodafone Zoozoos | Telecommunication |
| 6 | Lady Gaga | Musician band | -6 | Syracuse Orange | School sports team |
| 7 | Fox News | Media news publishing | -7 | St John's Red Storm | School sports team |
| 8 | Coca-Cola | Food beverages | -8 | 50 Cent | Musician band |
| 9 | ESPN | TV network | -9 | Max Bupa | Product service |
| 10 | LA Lakers | Professional sports team | -10 | Microsoft Developer | Computer technology |




(a) Most Influential 100 Brands

(b) Least Influential 100 Brands

**Figure 10. Brand Category Analysis**

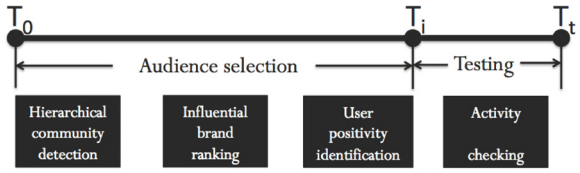



**Figure 11. Evaluation Framework**

and testing $<T_i, T_t>$. During the selection phase, we run our audience selection algorithm for a focal brand $F$ to obtain a target user pool $P$ of size $n$, where none of the users had activities on $F$ during the time period $<T_0, T_i>$. In the testing phase, we calculate the number of users out of these $n$ users who demonstrate positive activities (making positive comments or liking posts) on $F$.

To conduct performance comparisons, we select our baselines ($BL_1$, $BL_2$) very carefully. In addition, we also want to evaluate the effectiveness of individual components ($P_1$, $P_2$, $P_3$) within our framework and their different combinations. To the best of our knowledge, this is one of the first studies on audience selection using a social media platform such as Facebook. Although there is some related work as discussed earlier, they use different platforms where data are anonymous and inaccessible to us. Thus, there is no way for us to compare with these previous techniques. We list different targeting methods as follows:

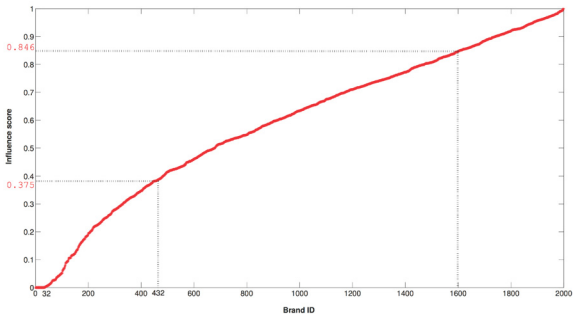



**Figure 12. Distribution of Brand Ranking Scores**

- For ($BL_1$), we randomly choose target users across all brands.
- For ($BL_2$), we first select top $k$ target brands ($b_1$, $b_2$, …, $b_k$) which have high numbers of common users with the focal brand. Then we randomly choose users from $b_1$, $b_2$, …, $b_k$.
- For ($P_1$), we first conduct hierarchical community detection and select brands ($b_1$, $b_2$, …, $b_k$) from the community which has the focal brand. Then we randomly choose users from $b_1$, $b_2$, …, $b_k$.
- For ($P_2$), we first select top $k$ target brands ($b_1$, $b_2$, …, $b_k$) based on global importance scores $bRank(b_i)$ calculated from our ranking algorithm. Then we randomly choose users from $b_1$, $b_2$, …, $b_k$.
- For ($P_3$), we choose users with high positivity based on sentiments expressed in their historical comments.
- For ($P_1 + P_2$), we first conduct hierarchical community detection to select $k$ target brands from the community which has the focal brand. We then select top (i.e., most important) brands from these target brands. Audience/users are chosen randomly from these important target brands.
- For ($P_1 + P_3$), we first conduct hierarchical community detection to select $k$ target brands from the community which has the focal brand. Then we choose users from these brands based on their sentiment positivity score.
- For ($P_2 + P_3$), we first select target brands based on brand ranking scores. Then users are chosen from these brands based on their positivity score.
- ($P_1 + P_2 + P_3$) is our audience selection framework incorporating all three components.

In our experiments, the selection period is from January 1, 2009, through May 1, 2012.

The testing period is from May 1, 2012, through January 1, 2013. We chose a sample of 10 focal brands ($F_1$, $F_2$, …, $F_{10}$) from the brands with highest 20% ranking scores and brands with lowest 20% ranking scores, respectively. Figure 12 shows that the threshold scores for the top 20% important brands and the bottom 20% less important brands are 0.846 and 0.375, respectively. We excluded 42 isolated nodes with the ranking score of 0 among the low ranking brands. We target $m = 100{,}000$ users for each focal brand $F_i$. We have the ability to adjust $k$ (i.e., the number of brands within the same community as the focal brand) to obtain $m$ if necessary (this may be required if the brands within a community do not together have a minimum of $m$ users). During the testing period, we calculate the number of users ($N_i$) who have activities on the focal brand $F_i$. Variables and their values are summarized in Table 5.

Tables 6 and 7 show the performance comparison of individual components of our framework and their combination, together with the two baselines. Each integer in the tables gives the number of users having activities on the focal brand in the testing period, and who did not have any focal brand activity in the training period. $P_1 + P_2 + P_3$ is our overall audience selection strategy. Table 6 and Table 7 show that our strategy achieves better performance, of up to 152 times increase when compared with $BL_1$, regardless of brands with high or low ranking scores. Among the three individual components in our approach, selecting users based on common community with the focal brand gives the best performance, followed by selection based on global importance scores and then by sentiment scores. This highlights the value of leveraging localized network information around a focal brand, as with community detection. The second baseline, which selects users based on brands that share high numbers of common users with the focal brand, also utilizes local information,

**Table 5. Evaluation Variables**

| Parameter | Meaning | Value Used in this Paper |
|---|---|---|
| $b$ | Number of focal brands | 10 |
| $k$ | Number of brands similar to the focal brand | Varied |
| $m$ | Number of target users | 100,000 |
| $F_i$ | The $i^{th}$ focal brand | $F_1, F_2, \ldots, F_{10}$ |
| $N_i$ | Number of users having positive activities on $F_i$ in the testing period | Varied |

**Table 6. Performance Comparison for High-Ranking Brands**

| Brand Name | $BL_1$ | $BL_2$ | $P_1$ | $P_2$ | $P_3$ | $P_1 + P_2$ | $P_1 + P_3$ | $P_2 + P_3$ | $P_1 + P_2 + P_3$ |
|---|---|---|---|---|---|---|---|---|---|
| Pepsi | 30 | 211 | 278 | 250 | 76 | 506 | 336 | 299 | 1,247 |
| Nokia | 28 | 261 | 478 | 162 | 44 | 735 | 620 | 207 | 1,364 |
| Mitt Romney | 25 | 1,008 | 1,149 | 792 | 324 | 1,886 | 1,323 | 1,056 | 3,821 |
| Cristiano Ronaldo | 102 | 2,034 | 1,795 | 886 | 277 | 2,223 | 1,810 | 1,106 | 2,419 |
| McDonald's | 70 | 1,373 | 1,720 | 1,006 | 218 | 1,889 | 1,405 | 1,104 | 2,014 |
| Starbucks | 63 | 1,508 | 1,043 | 792 | 173 | 1,695 | 1,620 | 1,015 | 2,107 |
| Xbox | 60 | 300 | 407 | 331 | 107 | 1,044 | 770 | 392 | 1,521 |
| Tide | 56 | 303 | 448 | 390 | 104 | 975 | 822 | 531 | 1,218 |
| Chicago Bulls | 23 | 204 | 318 | 217 | 37 | 550 | 372 | 284 | 1,249 |
| Windows Phone | 81 | 1,502 | 1,344 | 570 | 173 | 1,444 | 1,367 | 589 | 1,566 |

**Table 7. Performance Comparison for Low-Ranking Brands**

| Brand Name | $BL_1$ | $BL_2$ | $P_1$ | $P_2$ | $P_3$ | $P_1 + P_2$ | $P_1 + P_3$ | $P_2 + P_3$ | $P_1 + P_2 + P_3$ |
|---|---|---|---|---|---|---|---|---|---|
| Vizio | 38 | 111 | 203 | 100 | 41 | 218 | 204 | 102 | 431 |
| BMW USA | 36 | 282 | 414 | 263 | 55 | 653 | 458 | 321 | 1,401 |
| Videograss | 18 | 95 | 126 | 78 | 45 | 177 | 132 | 90 | 225 |
| Umbro | 58 | 420 | 464 | 271 | 82 | 612 | 457 | 390 | 739 |
| Staples | 21 | 367 | 483 | 298 | 77 | 672 | 505 | 114 | 801 |
| Baby Phat | 47 | 204 | 310 | 235 | 32 | 348 | 324 | 237 | 506 |
| California Academy of Science | 18 | 124 | 202 | 117 | 19 | 260 | 193 | 120 | 271 |
| Lands End | 27 | 251 | 683 | 275 | 150 | 769 | 704 | 166 | 1,604 |
| Alternative Press | 36 | 120 | 197 | 94 | 22 | 216 | 202 | 107 | 417 |
| Zegna | 13 | 78 | 102 | 67 | 14 | 118 | 105 | 71 | 153 |

and results in much higher numbers of test period users for the focal brand than with random selection as in the first baseline. Community detection is seen to yield higher performance than the second baseline on most brands. Sentiment alone is inadequate for identifying target users.

## *Further Analysis and Results*

Each brand has some properties, such as the size (the number of fans) and the importance (ranking score). To help examine relationships between these properties and different targeting methods, we first define three metrics, two conversion lifts and one conversion rate:

- lift from the first baseline $\left(CL_1 = \frac{N_1 \text{ under } P_1+P_2+P_3}{N_i \text{ under } BL_1}\right)$
- lift from the second baseline $\left(CL_2 = \frac{N_i \text{ under } P_1+P_2+P_3}{N_i \text{ under } BL_2}\right)$
- conversion rate based on the number of users we target $\left(CR = \frac{N_i}{100{,}000}\right)$

**Table 8. Lift and Conversion Rates for High-Ranking Brands**

| Brand Name | Brand Size | Ranking Score | $CL_1$ | $CL_2$ | CR (%) |
|---|---|---|---|---|---|
| Pepsi | 9,000,230 | 0.997 | 41.57 | 5.91 | 1.25 |
| Nokia | 8,345,543 | 0.972 | 48.71 | 5.23 | 1.36 |
| Mitt Romney | 7,546,493 | 0.900 | 152.84 | 3.79 | 3.82 |
| Cristiano Ronaldo | 46,121,055 | 0.947 | 23.72 | 1.19 | 2.42 |
| Mcdonald's | 22,923,442 | 0.962 | 28.77 | 1.47 | 2.01 |
| Starbucks | 30,723,302 | 0.961 | 33.44 | 1.40 | 2.10 |
| Xbox | 1,809,066 | 1.000 | 25.35 | 5.07 | 1.52 |
| Tide | 3,133,344 | 0.967 | 21.75 | 4.02 | 1.22 |
| Chicago Bulls | 702,904 | 0.851 | 54.30 | 6.12 | 1.25 |
| Windows Phone | 1,339,704 | 0.960 | 19.33 | 1.04 | 1.57 |

**Table 9. Lift and Conversion Rates for Low-Ranking Brands**

| Brand Name | Brand Size | Ranking Score | $CL_1$ | $CL_2$ | CR (%) |
|---|---|---|---|---|---|
| Vizio | 121,089 | 0.001 | 11.34 | 3.88 | 0.43 |
| BMW USA | 772,563 | 0.095 | 38.92 | 4.97 | 1.40 |
| Videograss | 12,136 | 0.001 | 12.50 | 2.39 | 0.23 |
| Umbro | 353,593 | 0.001 | 12.74 | 1.76 | 0.74 |
| Staples | 506,887 | 0.144 | 38.14 | 2.18 | 0.80 |
| Baby Phat | 185,154 | 0.001 | 10.77 | 2.48 | 0.51 |
| California Academy of Science | 53,561 | 0.287 | 15.06 | 2.19 | 0.27 |
| Lands End | 947,109 | 0.217 | 59.41 | 6.39 | 1.60 |
| Alternative Press | 121,097 | 0.001 | 11.58 | 3.48 | 0.42 |
| Zegna | 88,846 | 0.093 | 11.77 | 1.96 | 0.15 |

Table 8 and Table 9 show corresponding statistics for brands with high-ranking scores and low-ranking scores, respectively.

These three metrics are critical for demonstrating the performance of our technique for audience selection. However, there are some additional challenges that need to be addressed in assessing real performance. The first one is to show that the users we select will actually click on the ads. The second is to guarantee that the increase in activity level of targeted users on the focal brand is not caused by other special events on Facebook, such as brand promotions or discounts and friend referrals. These can only be answered by placing real ads and controlling for special events pertaining to the focal brands. For the purpose of evaluation here, we examine the effectiveness of our strategy in identifying target users who will display high user engagement with the focal brands; the evaluation is based on multiple brands, with varying brand ranking scores and size.

Tables 8 and 9 tell us that regardless of the brand ranking scores, our framework results in significantly high lifts and conversion rates. They do not have any correlation with brand ranking scores. The results in Tables 8 and 9 also show generally lower conversion rates with brands having low ranking scores. This relates to the lower brand size for most of the low ranking brands in Table 9; where brand size reaches levels comparable with brands in Table 8, the conversion rates are also higher. Given that conversion rate is determined as the number of users converted to engage with the focal brand, as a ratio of the 100,000 users targeted, even the seemingly low values in Table 9 (for example, 0.23 for Videograss) reflect significant improvements over the baselines (12.5 times better than random user selection and 2.39 times better than selecting users based on brands that share the most users with this focal brand).

## Conclusions and Future Work

This paper presents a framework to analyze large-scale data on user historical activities from a social media platform to identify audiences for online advertising. Audience selection is based on implicit brand–brand networks obtained from user

activities in brand communities. To our knowledge it is one of the first studies that develops networks showing relationships between brands based on a large social media dataset.

Our targeting framework includes four phases: extracting and normalizing brand–brand networks, finding a set of closely related brands with respect to a focal brand, identifying a subset of influential brands, and selecting target audiences from selected brands. We design a novel evaluation approach based on available data to test the effectiveness of our targeting framework as compared to baselines. The experiments show that our approach results in significant performance improvement as compared to baselines.

Our research provides a way to increase user social engagement for a focal brand. As shown in previous work, increased user engagement leads to increased purchase and revenues (Oestreicher-Singer and Zalmanson 2013). Experiments in this paper have shown that users identified by our framework are potentially interested in interacting with focal brands; such users can be selected as targets for advertising to increase engagement and loyalty. In addition, the brand–brand network reveals relationships between online brand communities and provides useful insights for brand managers. It can be used to obtain a deep understanding of customer interests and brand engagement, and how these evolve over time. Closely related brands, identified through community detection on the brand–brand network, can also form the basis for customer segmentation. The relationship between a brand and its consumers is an active area of research (Fournier, 1998), and the literature on brand communities (Fournier and Lee 2009; Muniz and O'Guinn 2001) has examined consumers with common interests in a brand. The implicit brand–brand networks developed in our work reveal a brand's affinity to other brands from a customer's point of view, and will be useful for extending the research on consumer brand relationships to consider consumers' engagement across brands.

Our current research does not consider the content of posts made by a brand. We analyze the content of comments to identify user positivity based on their sentiment of comments across all brands. Analyzing user sentiment from comments is especially important since all other social actions (e.g., like, share, etc.) are positively inclined. Considering user sentiment specific to individual brands can provide additional useful information for discerning customers based on their extent of positivity/negativity toward a brand. Incorporating users' brand-specific sentiments for audience targeting is a topic for future research. Relating the content of posts to user sentiment provides further avenues for fine-grained analyses.

As noted earlier, the brand–brand networks provide a unique view of relationships between brands from the perspective of consumers' overlapping interests. They provide a basis for audience selection, but have not been the focus of detailed analysis in this study. Such networks can be of significant interest for exploring brand interrelationships and the nature of consumer activities across related brand communities for various marketing purposes. In other published work, we described many interesting properties of brand–brand networks, for example, brand influence score has negative correlation with brand sentiments and positive correlation with the size of the brand, brands within close geographic proximity are more likely to be in the same community, and influential brands are likely to be in the same community (Zhang et al. 2014). The present study develops an implicit brand–brand network from historical data on user activity over multiple years. With changes in brand promotions and evolving user preferences over time, it would be interesting to consider changes in brand networks created by changes in user activity. In our future work, we will conduct dynamic network analysis, by extracting multiple implicit brand–brand networks using time windows (e.g., a year), to examine evolving properties of networks and their effects on audience targeting.

Our evaluation method helps demonstrate the value of the proposed framework for audience selection, compared with baselines and other targeting methods, and demonstrates significant performance improvements. Field studies using controlled tests can help further assess the effectiveness of the proposed framework.

## References


Aral, S., and Walker, D. 2011. "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management Science* (57:9), pp. 1623-639.

Baesens, B. 2014. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, Hoboken, NJ: Wiley.

Bakshy, E., Eckles, D., Yan, R., and Rosenn, I. 2012. "Social Influence in Social Advertising: Evidence from Field Experiments," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, New York: ACM, pp. 146-161.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* (489:7415), pp. 295-298.

Brodie, R. J., Llic, A., Juric, B., and Hollebeek, L. 2013. "Consumer Engagement in a Virtual Brand Community: An Exploratory Analysis," *Journal of Business Research* (66:1), pp. 105-114.

Bruyn, A. D., and Lilien, G. L. 2008. "A Multi-Stage Model of Word-of-Mouth Influence Through Viral Marketing," *International Journal of Research in Marketing* (25:3), pp. 151-163.

Centola, D. 2010. "The Spread of Behavior in an Online Social Network Experiment," *Science* (329:5996), pp. 1194-1197.

Chau, M., and Xu, J. 2012. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), pp. 1189-1216.

De Valck, K., van Bruggen, G. H., and Wierenga, B. 2009. "Virtual Communities: A Marketing Perspective," *Decision Support Systems* (47:3), pp. 185-203.

De Vries, L., Gensler, S., and Leeflang, P. S. H. 2012. "Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing," *Journal of Interactive Marketing* (26:2), pp. 83-91.

Domingos, P., and Richardson, M. 2001. "Mining the Network Value of Customers," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 57-66.

Fournier, S. 1998. "Consumers and Their Brands: Developing Relationship Theory in Consumer Research," *Journal of Consumer Research* (24:4), pp. 343-373.

Fournier, S., and Lee, L. 2009. "Getting Brand Communities Right," *Harvard Business Review* (87:4), pp. 105-111.

Godes, D., and Mayzlin, D. 2009. "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science* (28:4), pp. 721-739.

Goel, S., and Goldstein, D. G. 2013. "Predicting Individual Behavior with Social Networks," *Marketing Science* (33:1), pp. 82-93.

Goldfarb, A., and Tucker, C. 2011. "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science* (30:3), pp. 389-404.

Gopal, R., Marsden, J. R., and Vanthienen, J. 2011. "Information Mining—Reflections on Recent Advancements and the Road Ahead in Data, Text, and Media Mining," *Decision Support Systems* (51:4), pp. 727-731.

Gruzd, A., and Wellman, B. 2014. "Networked Influence in Social Media: Introduction to the Special Issue," *American Behavioral Scientist* (58:10), pp. 1251-1259.

Hill, S., Provost, F., and Volinsky, C. 2006. "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science* (21:2), pp. 256-276.

Kwon, K. H., Stefanone, M. A., and Barnett, G. A. 2014. "Social Network Influence on Online Behavioral Choices: Exploring Group Formation on Social Network Sites," *American Behavioral Scientist* (58:10), pp. 1345-1360.

Lee, D., Hosanagar, K., and Nair, H. 2014. "The Effect of Advertising Content on Consumer Engagement: Evidence from Facebook " available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2290802.

Manchanda, P., Xie, Y., and Youn, N. 2008. "The Role of Targeted Communication and Contagion in Product Adoption," *Marketing Science* (27:6), pp. 961-976.

Mangold, W. G., and Faulds, D. J. 2009. "Social Media: The New Hybrid Element of the Promotion Mix," *Business Horizons* (52:4), pp. 357-365.

Marsden, J. R. 2008. "The Internet and DSS—Massive, Real-Time Data Availability is Changing the DSS Landscape," *Information Systems and e-Business Management* (6:2), pp. 193-203.

Muniz, A. M., and O'Guinn, T. C. 2001. "Brand Community," *Journal of Consumer Research* (27:4), pp. 412-432.

Newman, M. 2006. "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences of the USA* (103:23), pp. 8577-8582.

Oestreicher-Singer, G., and Zalmanson, L. 2013. "Content or Community? A Digital Business Strategy for Content Providers in the Social Age," *MIS Quarterly* (37:2), pp. 591-616.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing–Volume 10*, Stroudsburg, PA: Association for Computational Linguistics, pp. 79-86.

Provost, F., Dalessandro, B., Hook, R., Zhang, X., and Murray, A. 2009. "Audience Selection for On-Line Brand Advertising: Privacy-Friendly Social Network Targeting," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* New York: ACM, pp. 707-716.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. 2011. "User-Level Sentiment Analysis Incorporating Social Networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 1397-1405.

Van den Bulte, C. 2010. "Opportunities and Challenges in Studying Customer Networks," *The Connected Customer: The Changing Nature of Consumer and Business Markets*, S. H. K. Wyuts, M. G. Dekimpe, E. Gijsbrechts, and F. G. M. Pieters (eds.), New York: Routledge, pp. 7-35.

Watts, D. J., and Dodds, P. S. 2007. "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research* (34:4), pp. 441-458.

Zhang, K., Bhattacharyya, S., and Ram, S. 2014. "Empirical Analysis of Implicit Brand Networks on Social Media," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, New York: ACM, pp. 190-199.

Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D., Lee, K., Liao, W., and Choudhary, A. 2011. "SES: Sentiment Elicitation System for Social Media Data," in *Proceedings of the 11th International Conference on Data Mining Workshops*, Washington, DC: IEEE Computer Society, pp. 129-136.

## About the Authors

**Kunpeng Zhang** is an assistant professor of Decision, Operations & Information Technologies in the Robert H. Smith School of Business at the University of Maryland, College Park. Kunpeng received his Ph.D. from Northwestern University in 2013. His research is in the areas of big data analytics, social network analysis, machine learning, and natural language processing. He develops distributed techniques to analyze business, social media, and healthcare data. He has published articles in various journals, including *Neural Networks, Journal of Medical Internet Research,* and *Physical Review B*, and in computer science conference proceedings. He teaches courses on big data analytics and basic business programming.

**Siddhartha Bhattacharyya** is professor of Information and Decision Sciences in the College of Business Administration at the University of Illinois, Chicago. He received his Ph.D. from the

University of Florida in 1993. His research addresses data intensive applications, information management, agent-based models and evolutionary computation. His work appears in journals and conference proceedings including *Complex Systems*, *Decision Support Systems*, *Decision Sciences*, *Evolutionary Computation*, *IEEE Transactions on Evolutionary Computation*, *INFORMS Journal of Computing*, *Journal of Economic Dynamics and Control*, and *Social Networks.* He teaches courses in business data mining and technology project management and has received multiple teaching awards.

**Sudha Ram** is Anheuser-Busch Endowed Professor of MIS, and Entrepreneurship & Innovation in the Eller College of Management at the University of Arizona. She has joint faculty appointment as a professor of Computer Science. She is the director of the Advanced Database Research Group and codirector of INSITE: Center for Business Intelligence and Analytics (www.insiteua.org) at the University of Arizona. Sudha received a Ph.D. from the University of Illinois at Urbana-Champaign in 1985. Her research is in the areas of enterprise data management, business intelligence, large-scale networks, and data analytics. Her work uses different methods such as machine learning, statistical approaches, ontologies and conceptual modeling. She has published articles in such journals as *Communications of the ACM, IEEE Intelligent Systems, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Information Systems Research, Management Science,* and *MIS Quarterly*. Her research has been highlighted in several media outlets including NPR news. She was a speaker for a TED talk in December 2013 on "Creating a Smarter World with Big Data."

MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

# LARGE-SCALE NETWORK ANALYSIS FOR ONLINE SOCIAL BRAND ADVERTISING

**Kunpeng Zhang**
Department of DOIT, Robert H. Smith School of Business, University of Maryland,
College Park, MD 27042 U.S.A. {kzhang@rhsmith.umd.edu}

**Siddhartha Bhattacharyya**
Department of IDS, College of Business Administration, University of Illinois, Chicago,
Chicago, IL 60607 U.S.A. {sidb@uic.edu}

**Sudha Ram**
Department of MIS, Eller College of Management, University of Arizona,
Tucson, AZ 85721 U.S.A. {ram@eller.arizona.edu}

# Appendix A

## Network Generation

### *Algorithm 1: Chaining Two MapReduce Jobs to the Brand–Brand Network*

**Input**: A text file contains lines of <$brand_{id}$, $user_{id}$, # of activities>
**Output**: A text file contains lines of <$brand_i$, $brand_j$, # of common users>

1: /* The first job */
2: input: <$brand_{id}$, $user_{id}$> // Each line in the text file

3: **function** MAPPER
4: output < $user_{id}$, $brand_{id}$>
5: **end function**

6: **function** REDUCER
7: **for all** $v \in values$ **do**
8: add $v \rightarrow list$
9: **end for**
10: **for all** <$b$, $b_j$>, $b_i$, $b_j \in list$ **do**
11: add $v \rightarrow list$
12: **end for**
13: **output** <$k_2$, $v_2$>
14: **end function**

15: /* The second job */

```
16:  function IDENTITY MAPPER
17:  end function
18:  function REDUCER
19:        for all v ∈ values do
20:              sum += v
21:        end for
22:   output <key, sum>
23:  end function
```

# Appendix B

## Hierarchical Community Detection

### *Algorithm 2: Hierarchical Community Detection*

```
1:  C* ← {∅}
2:  function DIVIDE(B_n, s)    // s is the threshold and B_n is the network
3:     C: {C_1, C_2, …, C_k} ← Modularity-Based Detection (B_n)
4:     for all C_i ∈ C do     // this can be processed in parallel
5:           if | C_i | ≥ s then
6:                 C ← DIVIDE(C_i, s)
7:           else
8:                 C* ← C* ∪ C_i
9:           end if
10:    end for
11:    return C*
```

# Appendix C

## Brand Ranking

### *Algorithm 3: Distributed bRank: Mapper and Reducer Functions to Rank Brands*

1: /* The job for Mapper is to invert the input */

2: **function** MAPPER
3: **for all** $brand_j \in (brand_1, brand_2, \ldots, brand_k)$ **do**
4: **output** $brand_j \leftarrow <brand_i, rank_i * \frac{w_{ij}}{\sum w_i}>$ //$w_i$ is weights of all out-links from $i$
5: **end for**
6: **output** $brand_i \rightarrow brand_1, brand_2, \ldots, brand_k$
7: **end function**

8: /* The job for Reducer is to update the ranking using the in-links */

9: **function** REDUCER
10: Input is in a format of (∗). The key: $brand_k$
11: **for all** in-link $brand_i \in (brand_1, brand_2, \ldots, brand_n)$ **do**
12: $rank_k += rank_k * \frac{w_{ij}}{\sum w_i} \beta$ // is weights of all out-links from $i$
13: **end for**
14: $rank_k = (1 - \beta + rank_k) * C_n(k)$
15: **output** $<brand_k, rank_k> \rightarrow <brand_1, brand_2, \ldots, brand_n>$
// $brand_1, brand_2, \ldots, brand_n$ are out-links of $brand_k$
16: **end function**

After map function, we have temporary files in the following structure (∗):
$brand_k \rightarrow <brand_1, rank_1>$,
$<brand_2, rank_2>$,
…,
$<brand_n, rankn2>$,
$<brand_{k1}, brand_{k2}, \ldots, brand_{kn}>$

Where $brand_1, brand_2, \ldots, brand_n$ are in-links of $brand_n$ and $brand_{k1}, brand_{k2}, \ldots, brand_{kn}$ are out-links.